Published in final edited form as:

J Acoust Soc Am. 2006 January ; 119(1): 566–574.

Perceptual similarity of regional dialects of American English

Cynthia G. Clopper^{a)}, Susannah V. Levi, and David B. Pisoni

Department of Psychological and Brain Sciences, Indiana University, Bloomington, Indiana 47405

Abstract

Previous research on the perception of dialect variation has measured the perceptual similarity of talkers based on regional dialect using only indirect methods. In the present study, a paired comparison similarity ratings task was used to obtain direct measures of perceptual similarity. Naive listeners were asked to make explicit judgments about the similarity of a set of talkers based on regional dialect. The talkers represented four regional varieties of American English and both genders. Results revealed an additive effect of gender and dialect on mean similarity ratings and two primary dimensions of perceptual dialect similarity: geography (northern versus southern varieties) and dialect markedness (many versus few characteristic properties). The present findings are consistent with earlier research on the perception of dialect variation, as well as recent speech perception studies which demonstrate the integral role of talker gender in speech perception.

I. INTRODUCTION

A growing body of research has shown the importance of indexical properties in spoken language processing. Abercrombie (1967) described indexical properties of speech as those aspects of the signal which provide an index about the talker as an individual and as a member of social groups based on gender, ethnicity, and regional background. Research on the perception of the indexical properties of speech has typically involved one of two approaches. The first approach is to explore the interaction between indexical and linguistic properties in speech processing tasks, such as word recognition (Goldinger, 1996; Nygaard *et al.*, 1994) or sentence intelligibility (Nygaard and Pisoni, 1998). The second is to ask naive listeners to make explicit judgments about talkers' identity and social group membership, including judgments of talker similarity based on voice quality, gender, and regional dialect. The current study employed the latter method to directly assess naive listeners' perception of indexical properties of speech.

In one early study of the perception of talker similarity, Walden *et al.* (1978) asked naive listeners to rate the similarity of pairs of male talkers based on single word utterances using a seven-point similarity scale. A multidimensional scaling analysis of the results revealed two primary dimensions of within-gender talker similarity, word duration and F0. More recently, Kreiman and her colleagues (Kreiman *et al.*, 1994) asked trained linguists and speech-language pathologists to listen to pairs of male talkers with voice disorders and rate the relative breathiness and roughness of each pair using seven-point dissimilarity scales. Using multidimensional scaling techniques, they found that breathiness and roughness were the two primary dimensions of similarity for both the breathiness and the roughness ratings, suggesting that these two factors are integral components of judgments of disordered voices. More recently, Remez and his colleagues (Remez *et al.*, 2004) examined the role of dialect mismatch on the perceptual similarity of a set of female talkers. Listeners in Brooklyn, NY,

and Bloomington, IN, were asked to rate the similarity of two sets of female talkers. One group of talkers was native to Brooklyn and the other group was native to Bloomington. The authors found that the listeners' region of origin had an effect on the perception of talker similarity. However, they did not observe any strong acoustic correlates of perceptual talker similarity for the two listener groups.

All three of these studies exploring talker similarity used within-gender comparisons. However, gender appears to be a very salient indexical property for naive listeners. Lass *et al.* (1976) reported that gender identification performance based on sustained isolated vowels was very good in normal, whispered, and filtered conditions. Therefore, the limited design of the previous studies may have been warranted to avoid interactions between gender and voice quality in assigning similarity ratings.

Perceptual ratings directly related to gender similarity have also been obtained in several recent studies which explored the acoustic correlates of "masculine-sounding" speech. In one of these studies, Gaudio (1994) asked naive listeners to rate gay and straight male talkers on a seven-point gay-straight scale using a single-interval task. He found that participants were highly accurate in their judgments across several different reading styles. Using a two-interval task, Avery and Liss (1996) asked naive listeners to make relative masculinity judgments with a single talker as the reference throughout the experiment. The authors found that the relative masculinity judgments were correlated with several acoustic properties of the speech samples, including F0. Taken together, these studies suggest that F0 may also play a role in explicit judgments of gender similarity across a set of speakers with a range of gender prototypicality [see also Strand (1999)].

Unlike voice quality and gender, however, the perception of dialect similarity has not received much attention in speech perception research. Clopper (2004) indirectly examined dialect similarity using an auditory free classification task. Naive listeners were asked to group unfamiliar talkers by dialect similarity using sentence-length speech samples. The listeners were permitted to make as many groups as they wanted with as many talkers in each group as they wanted. Multidimensional scaling analyses revealed three primary dimensions of similarity that were consistent across two sets of stimulus materials: geography (northern versus southern), markedness (varieties with many distinguishing features versus varieties with few characteristic properties), and gender.

Sociolinguists and dialectologists have also studied dialect similarity, using computational and survey methods. For example, Nerbonne and his colleagues (Nerbonne and Heeringa, 2001; Nerbonne *et al.*, 1996) used string comparison metrics, such as Levenshtein distances, to compute dialect similarity scores for more than 100 Dutch dialects. In Japan, Mase (1999) used a map task to obtain judgments of dialect similarity from naive participants. He showed his participants a map of Japan and asked them to indicate where people spoke the same as them, where they spoke a little differently, where they spoke quite differently, and where they spoke an unintelligibly different variety. In general, his results were comparable to the more traditional dialect boundaries drawn by Japanese dialectologists.

The purpose of the current set of experiments was to explicitly examine perceptual dialect similarity using a two-interval similarity ratings task with a set of talkers from several different regional varieties of American English. Previous research on the perception of dialect variation by Clopper and her colleagues (Clopper, 2004; Clopper *et al.*, 2005a) has shown that naive listeners are able to accurately classify unfamiliar talkers with above chance performance in six-alternative forced-choice categorization tasks.

Experiment 1 was a replication of these earlier studies using a smaller set of talkers with fewer response alternatives. The reduction in response alternatives was motivated by results

of our earlier perception research and by production work in sociolinguistics (e.g., Labov, 1998). A clustering analysis of Clopper's (2004) free classification data suggested that naive listeners have four primary perceptual categories for regional varieties of American English: Northeast, North, South, and Midland/West. The Northeast dialect included Mid-Atlantic talkers and r-less New England talkers. New England talkers who were not r-less were perceptually similar to Midland and Western talkers. In our previous work (Clopper, 2004; Clopper *et al.*, 2005a), we interpreted these perceptual similarity findings in terms of the markedness of the different varieties. As noted above, markedness refers to the extent to which a given dialect has unique characteristics that differentiate it from the other dialects of American English. In this sense, Northern, Southern, Mid-Atlantic, and r-less New England varieties are "marked" because they have many features that distinguish them from other dialects. However, Midland, Western, and r-ful New England varieties are considered less marked because they have fewer characteristic properties.

These perceptual results are also consistent with Labov's (1998) division of American English into three broad dialect categories (North, South, and the "Third Dialect"), based on the vowel systems of the different varieties. The "Third Dialect" includes New England, the Midland, and the West; Mid-Atlantic is treated as an exception. Each of the other primary dialects is also composed of subdialects, such as the Inland North in the Northern dialect or the Coastal Southeast in the Southern dialect. In addition, intradialect variation exists between talkers in all regions, such that some talkers are more likely to produce the characteristic properties of their dialect than others. However, the regions defined by Labov (1998) provide a picture of regional variation of the United States painted with a broad stroke that is supported by the perception of dialect variation by naive listeners.

Based on the earlier perception and production research, we selected talkers from the following four broad regional varieties of American English for the current study: Mid-Atlantic, North, South, and "General American."¹ We predicted that naive listeners would perform somewhat better on the four-alternative categorization task than the previous six-alternative tasks because some of the more confusable dialects were eliminated as response alternatives and because chance performance was higher in the new task with fewer alternatives.

Experiment 2 used a direct two-interval dialect similarity ratings task with the same set of talkers as in experiment 1. We expected that naive listeners would assign overall higher similarity ratings to pairs of talkers from the same dialect region than to pairs of talkers from different dialect regions and that the patterns of perceptual similarity would reflect the primary dimensions of dialect similarity reported in the earlier free classification experiment by Clopper (2004), including geography, markedness, and gender.

II. EXPERIMENT 1

A. Methods

1. Listeners—Thirty-seven Indiana University undergraduates participated as listeners in experiment 1. Prior to data analysis, data from nine participants were excluded: two were bilingual, two had one or more parents whose first language was not English, two had a history of a hearing or speech disorder, one did not complete the entire experiment, one was

¹We have chosen to refer to this fourth dialect as "General American" for several reasons. First, in experiment 1, we needed to provide a label to our listeners for this region and "General American" was both nontechnical and intuitive for our listeners. Second, while we recognize that this term can imply a national norm or standard, we have not encountered an alternative that avoids this problem while maintaining descriptive adequacy. Thus, in this paper, "General American" is defined as the combination of r-less New England, Midland, and Western varieties and is comparable to Labov's (1998) "Third Dialect." It is not intended to carry any prestige, standard, or normative value.

significantly older than the other participants, and one performed statistically at chance across both blocks and all four groups of talkers. The remaining 28 listeners were all 18–25-year-old monolingual native speakers of American English with native English-speaking parents and no history of speech or hearing disorders reported at the time of testing. The residential history of the listeners varied, but the majority ($N=16$) had lived in the Midwest (Midland or North dialect region) for their entire lives. The remainder were either lifetime residents of another region ($N=2$) or had lived in multiple different dialect regions before attending Indiana University in Bloomington ($N=10$). The listeners received partial credit in a psychology course for participating.

2. Talkers—The talkers in experiment 1 were a subset of the talkers included in the Nationwide Speech Project (NSP) corpus (Clopper and Pisoni, unpublished). The NSP corpus includes recordings of five male and five female lifetime residents of six dialect regions in the United States: New England, Mid-Atlantic, North, Midland, South, and West. Thirty-two talkers were selected for the current study. In particular, four males and four females were randomly selected from each of the Mid-Atlantic, Northern, and Southern regions. The remaining four male and four female talkers comprised the set of “General American” talkers and were randomly drawn from the New England, Midland, and Western regions. Thus, the 32 talkers in experiment 1 represented both genders and four dialects of American English: Mid-Atlantic, North, South, and General American.

The Northern dialect is characterized by the Northern Cities Chain Shift (NCCS), which involves the clockwise rotation of the low and low-mid vowels, beginning with the fronting and raising of /æ/. Other aspects of the NCCS include backing of /ε/ and /ʌ/, fronting and lowering of /ɔ/, and fronting of /ɑ/ (Clopper *et al.*, 2005b; Labov, 1998). The Southern dialect is characterized by the fronting of the high back vowels /u/, /ʊ/, and /ow/, /ɑ y/ and /oy/ monophthongization, and the Southern Vowel Shift (SVS), which includes the peripheralization of the front lax vowels /ɪ/ and /ε/, and the centralization of the front tense vowels /i/ and /ey/ (Clopper *et al.*, 2005b; Labov, 1998; Thomas, 2001). The Mid-Atlantic dialect is traditionally characterized by the raising of /ɔ/ (Labov, 1972), but the Mid-Atlantic talkers included in the NSP corpus did not exhibit this particular vowel shift. Instead, they showed evidence of backing /ʌ/ and /ɑ/ and fronting /ɔ/ (Clopper *et al.*, 2005). General American, or “the Third Dialect,” is characterized by the merger of the low back vowels /ɑ/ and /ɔ/ (Labov, 1998). Several recent acoustic analyses also suggest that /u/ fronting is spreading to regions beyond the South, including the West and Midland regions (Clopper *et al.*, 2005b; Thomas, 2001). Finally, r-lessness is a prominent property of both New England and coastal Southern dialects, but none of the New England or Southern talkers in the NSP corpus exhibited r-lessness in their speech.

3. Stimulus materials—Two different sentences from each talker were selected for the stimulus materials in experiment 1, for a total of 64 different stimulus items. The sentences were short, meaningful English sentences taken from the Speech Perception in Noise test (Kalikow *et al.*, 1977). All of the sentences were “high probability” sentences, which means that the final word was semantically predictable from the preceding context. The sentences for each talker were selected to include dialect-specific vowel shifts, so that regional variation was present in the stimulus materials. For example, sentences for the Northern talkers frequently contained fronted /æ/ or raised /ɑ/ and sentences for the Southern talkers often contained monophthongal /ɑ y/ or fronted /u/. A complete list of the test sentences is included in the Appendix. No sentence was repeated during the course of the experiment.

4. Procedure—The listeners were seated in a quiet testing room at personal computers equipped with headphones (Beyerdynamic DT100) and a mouse. On each trial, the listeners heard a single sentence produced by one of the talkers and were asked to identify which

region of the United States the talker was from. The regions were presented on a color-coded labeled map of the United States, as shown in Fig. 1. The listeners were permitted to hear each sentence as many times as they wanted before making their response, but they did not receive any feedback about the accuracy of their responses.

The experiment was divided into two blocks. Each block contained one sentence produced by each talker, for a total of 32 trials per block. The sentences were randomly assigned to the first or second block for each listener and the trials were presented in random order within each block. The stimuli were presented over headphones at a comfortable listening level (approximately 70 dB SPL). The listeners made their responses by clicking on the labeled button for the region that they thought each talker was from. The experiment took approximately 15 min to complete.

B. Results

A summary of the listeners' categorization performance in experiment 1 is shown in Fig. 2. In a four-alternative task, chance performance is 25% and a binomial test confirmed that overall performance was statistically above chance ($p < 0.05$). The listeners' performance was above chance for all four talker groups.

A repeated measures ANOVA with talker dialect (Mid-Atlantic, North, South, or General American) and experimental block (first or second) as within-subject factors revealed a significant main effect of talker dialect [$F(3,81) = 11.3$, $p < 0.001$]. *Post hoc* paired comparison *t* tests revealed that performance on the General American talkers was significantly better than performance on the Mid-Atlantic, Northern, and Southern talkers (all $p < 0.001$). Performance on the Mid-Atlantic talkers was significantly better than performance on the Northern talkers ($p = 0.03$). Performance on the Southern talkers did not differ significantly from performance on either the Mid-Atlantic or the Northern talkers. The main effect of experimental block and the dialect $\times$ block interaction were not significant.

To assess the effects of talker gender on performance, a paired sample *t* test was conducted on the mean accuracy scores for male and female talkers, collapsed across talker dialect and experimental block. The male and female talkers were categorized with 40% and 43% accuracy, respectively. This difference was not significant [$t(27) = 1.08$, n.s.]. Thus, talker gender did not affect dialect categorization performance.

As expected, overall raw percent correct scores were higher on this task than on the earlier six-alternative forced-choice task using stimulus materials from the same speech corpus (Clopper, 2004). Performance cannot be compared directly using percent correct scores, however, because the different number of response alternatives in the two tasks altered the level of chance performance (17% in the six-alternative task versus 25% in the four-alternative task). In order to compare the previous results to those obtained in the current experiment, the data from each listener were converted into a relative information transmitted score (Miller, 1953; Miller and Nicely, 1955), using the equation in (1), where p_i is the probability of stimulus category i , p_j is the probability of response alternative j , and p_{ij} is the probability of response j given stimulus i . These information transmitted scores provide another index of performance accuracy, beyond raw percent correct, which normalizes the data for the number of response alternatives and the participants' response biases:

$$T(x, y) = \frac{-\sum p_{ij} \log_2 (p_i p_j / p_{ij})}{-\sum p_i \log_2 (p_i)}. \quad (1)$$

Information transmitted scores were also calculated from the earlier six-alternative data. The mean information transmitted scores were 14% (SD=7%) and 15% (SD =6%) in the current four-alternative categorization experiment and the earlier six-alternative task, respectively. An independent samples *t* test revealed no significant difference across the two tasks in terms of relative information transmitted [$t(125) = -0.77$, n.s.].

C. Discussion

Overall performance in the four-alternative forced-choice dialect categorization task was approximately 42% correct, which is significantly above chance performance (25%). While this overall level of accuracy is higher than that reported in the previous studies of dialect categorization using six alternatives (e.g., Clopper, 2004; Clopper *et al.*, 2005a), the information transmitted analysis revealed that performance across the two tasks was comparable when the different levels of chance performance were accounted for. Thus, explicit dialect categorization is a difficult task for naive listeners, regardless of whether they are presented with four or six response alternatives.

The finding that naive listeners were able to perform above chance on these tasks, however, suggests that they were able to reliably classify talkers from different regions of the United States. We would therefore also expect them to be able to make consistent similarity judgments between pairs of talkers in a paired comparison similarity ratings task. Experiment 2 was designed to obtain explicit direct judgments of regional dialect similarity for all possible pairwise comparisons of the set of 32 talkers used in experiment 1.

III. EXPERIMENT 2

A. Methods

1. Listeners—One hundred and seven Indiana University undergraduates participated as listeners in experiment 2. Prior to data analysis, data from 14 listeners were removed. Seven participants knew one or more of the talkers by name, two did not finish the experiment, two did not meet our residential history requirements, one reported a history of a hearing or speech disorder, one had one or more parents whose first language was not English, and one had previously participated in a related experiment during the same semester. The remaining 93 listeners were all 18–25-year-old monolingual native speakers of American English with no history of hearing or speech disorders reported at the time of testing. In addition, both parents of each listener were also native speakers of English. The listeners each received \$15 for participating.

The listeners were assigned to four different groups, based on their residential history. Twenty-four of the listeners, who had lived only in the Midland dialect region, were assigned to the Non-Mobile Midland group. Similarly, 24 listeners who had lived only in the Northern dialect region prior to attending school in Bloomington, IN, were assigned to the Non-Mobile North group. The remaining 45 listeners had lived in more than one dialect region prior to attending Indiana University. Of these 45 Mobile listeners, the 21 whose parents lived in the Northern dialect region at the time of testing were assigned to the Mobile North group and the 24 whose parents lived in the Midland dialect region at the time of testing were assigned to the Mobile Midland group.

2. Talkers—The same 32 talkers were used in experiment 2 as in experiment 1.

3. Stimulus materials—The same 64 stimulus items were used in experiment 2 as in experiment 1.

4. Procedure—The listeners were seated in a quiet testing room in front of personal computers equipped with headphones (Beyerdynamic DT100) and a mouse. On each trial, the listeners heard two sentences produced by two different talkers separated by 500 ms of silence. They were then asked to indicate how likely it was that the two talkers came from the same part of the United States using a scale from 1 (“Not at all likely”) to 7 (“Very likely”). The listeners were asked to make a direct and explicit dialect similarity judgment, with higher numbers reflecting greater dialect similarity than lower numbers. The listeners did not receive any feedback about their responses.

One sentence from each talker was randomly selected and assigned to List A. The remaining sentence from each talker was assigned to List B (see the Appendix). Each listener heard the stimulus items from either List A or List B, so that there was a perfect one-to-one relationship between sentence content and talker throughout the experiment for each listener. Half of the listeners in each listener group heard the List A sentences and half of the listeners in each listener group heard the List B sentences.

Each listener heard all 32 talkers paired with each of the other talkers one time, for a total of 496 trials. Of the 496 trials, 112 included two talkers from the same dialect region (“same-dialect” trials) and the remaining 384 trials included two talkers from different dialect regions (“different-dialect” trials). In addition, 240 trials included two talkers with the same gender (“same-gender” trials) and the remaining 256 trials included one male and one female talker (“different-gender” trials). The order of presentation of the talkers within each pair was determined randomly, as was the order of presentation of the trials. The experiment was divided into eight blocks, with a short break provided after every 62 trials. Experiment 2 took approximately 75 min to complete.

B. Results

Figure 3 shows the mean similarity ratings for same-dialect and different-dialect pairs for each of the four listener groups. Performance across the four listener groups was highly consistent and all four groups showed overall higher similarity ratings for same-dialect pairs than different-dialect pairs. An initial inspection of the data revealed that all but three (3%) of the listeners used the entire seven-point scale in making their judgments and that the responses were approximately normally distributed between 1 and 7. Therefore, no normalization was necessary and the analyses were conducted on the raw perceptual ratings scores.

A repeated measures ANOVA with dialect (same-dialect versus different-dialect) and block (first through eighth) as within-subject factors and stimulus list (List A versus List B) and listener group (Mobile North, Mobile Midland, Non-Mobile North, or Non-Mobile Midland) as between-subject factors revealed a significant main effect of dialect [$F(1,595) = 41.6, p < 0.001$]. The same-dialect pairs were rated higher than the different-dialect pairs. *Post hoc* paired comparison *t* tests revealed that this effect was highly significant for all four listener groups (all $p < 0.001$). The main effect of block was also significant [$F(7,595) = 2.7, p = 0.009$], but *post hoc* Tukey tests did not reveal any significant pairwise comparisons. The main effects of stimulus list and listener group were not significant.

Two of the interactions were also significant. First, the dialect $\times$ stimulus list interaction was significant [$F(1,595) = 17.4, p < 0.001$]. While the overall ratings did not differ between List A and List B, *post hoc t* tests revealed that different-dialect pairs were rated more similar for List A than List B stimuli (List A $M = 3.75, SD = 0.50$; List B $M = 3.42, SD = 0.59$; $p < 0.001$). Ratings for same-dialect pairs for List A and List B were not significantly different (List A $M = 3.83, SD = 0.59$; List B $M = 3.79, SD = 0.73$; n.s.). The second significant interaction was the block $\times$ stimulus list interaction [$F(7,595) = 3.7, p = 0.001$]. Once again, while the overall

ratings did not differ across the two stimulus lists, *post hoc t* tests revealed that overall ratings were higher for List A than List B in blocks 3–7 (all $p < 0.02$). Ratings did not differ across the two lists for blocks 1, 2, and 8. Taken together these two interactions suggest that the sentences included in List A may have been more difficult to discriminate based on dialect, although the primary findings are robust across both stimulus sets. None of the other two- or three-way interactions were significant.

A repeated measures ANOVA with gender (same-gender versus different-gender) and dialect (same-dialect versus different-dialect) as within-subject factors was conducted to assess the effects of talker gender in the similarity ratings task. The repeated measures ANOVA revealed significant main effects of gender [$F(1,92) = 39.7, p < 0.001$] and dialect [$F(1,92) = 23.0, p < 0.001$]. As in the previous analysis, same-dialect pairs were rated significantly higher than different-dialect pairs. In addition, same-gender pairs were rated significantly higher than different-gender pairs. The gender $\times$ dialect interaction was not significant.

To examine the similarity ratings in more detail and obtain measures of the underlying similarity space, a multidimensional scaling analysis was conducted. For each listener, a 32×32 talker similarity matrix was constructed by assigning the similarity rating assigned to each talker pair to one cell of the matrix. The 93 talker similarity matrices were then submitted to an INDSCAL analysis to explore the role of individual listener differences in the perceptual similarity of the talkers. Two-, three-, and four-dimensional solutions were obtained using INDSCAL with mean stress values of 0.37, 0.28, and 0.23, respectively. While these results suggest an “elbow” at the three-dimensional solution, the two-dimensional solution was selected for interpretation and discussion for two reasons. First, the two-dimensional solution was highly interpretable and the addition of the third dimension did not contribute meaningfully to the analysis. Second, the two-dimensional model required fewer free parameters and was therefore preferable as a more parsimonious analysis, despite its lower overall level of fit to the data.

The two dimensions of the multidimensional scaling solution are plotted in Fig. 4. Each symbol represents a different talker. The shapes represent the four different dialect groups (squares for Mid-Atlantic, triangles for North, circles for South, and diamonds for General American). Male talkers are indicated by the filled symbols and female talkers are represented by the open symbols. Dimension 1 separates the Northern and Mid-Atlantic talkers on the left and the Southern and General American talkers on the right. The first dimension can therefore be interpreted as distinguishing the North from the South (or non-North). The second dimension separates the Mid-Atlantic and Southern talkers at the bottom from the Northern and General American talkers on the top, which can be interpreted in terms of dialect markedness.² Marked dialects contain many characteristic properties that distinguish them from other dialects, whereas unmarked dialects contain fewer characteristic features. In Fig. 4, the perceptually marked dialects are at the bottom (Mid-Atlantic and Southern) and the perceptually less-marked dialects are at the top (Northern and General American). Thus, the two underlying dimensions extracted from the INDSCAL analysis can be described in terms of geography (north versus south) and markedness (marked versus unmarked).

²This dimension could also be interpreted as an East versus West dimension with the Southern and Mid-Atlantic dialects in the East and the Northern and General American dialects in the West. However, this interpretation seems to suggest that the Southern dialect is limited to the eastern United States when, in fact, it extends all the way across the southern part of the country into Texas. In addition, our interpretation of this dimension in terms of markedness is more consistent with our previous research on the perceptual similarity of dialect variation (e.g., Clopper, 2004).

The INDSCAL analysis also returned dimension weights for each of the 93 listeners. The mean normalized weights for each listener group and each stimulus list are shown in Table I. Two-way ANOVAs on the normalized subject weights for each dimension with listener group (Mobile North, Mobile Midland, Non-Mobile North, or Non-Mobile Midland) and stimulus list (List A versus List B) as the factors revealed a significant main effect of stimulus list for both dimensions [$F(1,92) = 14.9, p < 0.001$ for dimension 1 and $F(1,92) = 14.9, p < 0.001$ for dimension 2]. The listeners who heard stimulus materials from List B attended more to the first dimension than the listeners who heard stimulus materials from List A. The participants who heard List A, however, attended more to the second dimension than the participants who heard List B. Thus, the stimulus list interactions reported above may reflect different strategies employed by the listeners in the two groups. Linguistic markedness was more salient for the listeners who were responding to List A stimulus materials, whereas geography was more salient for the listeners who were responding to List B stimulus materials.

The main effect of listener group was not significant. However, an inspection of the mean normalized subject weights suggests that the Mobile listeners tended to weight both dimensions equally, whereas the Non-Mobile listeners attended more to the geographic dimension (dimension 1) than the markedness dimension (dimension 2). The Non-Mobile listeners attended to the geographic dimension more than the Mobile listeners and to the markedness dimension less than the Mobile listeners, although neither of these differences were significant by independent sample *t* tests. In general, the dimension weights were highly consistent across the listener groups, with more attention paid to geography than markedness.

C. Discussion

The results of the paired comparison similarity ratings task demonstrate that naive listeners' direct judgments about the perceptual similarity of the dialect of two talkers reflect the regional background of the talkers. The initial analysis of the mean ratings revealed that the listeners consistently rated pairs of talkers from the same dialect region as more similar than pairs of talkers from different regions. This finding was consistent across all four listener groups, all eight blocks of trials, and both sets of stimulus materials. It is important to emphasize here that the paired comparison task was designed such that all of the similarity judgments were made across two different sentences, which means that the listeners were required to judge dialect similarity in the absence of identical linguistic content.

Talker gender was also found to be an important component in the similarity judgments, however. In particular, same-gender pairs were consistently rated as more similar than different-gender pairs, regardless of whether they were same-dialect or different-dialect pairs. This result is a classic "additive" effect, which suggests that gender and dialect are independent factors that contribute to the judgment of dialect similarity (Sternberg, 1998). Unlike the earlier free classification study (Clopper, 2004), in which gender emerged as one of the primary dimensions of perceptual similarity in the multidimensional scaling analysis, gender was not an important dimension in the INDSCAL analysis of the similarity ratings data. Given that the listeners were not given any specific instructions regarding the role of gender in determining dialect similarity, however, the gender effect on the overall mean similarity ratings suggests that naive listeners are unable to selectively ignore talker gender when making judgments about regional dialect (see also Mullennix and Pisoni, 1990).

The multidimensional scaling analysis uncovered two dimensions of perceptual dialect similarity: geography (north versus south) and markedness. These perceptual dimensions create a space in which the Southern talkers are located in the Marked Southern quadrant (bottom right in Fig. 4), the General American talkers are located in the Unmarked

Southern/Non-Northern quadrant (top right in Fig. 4), the Northern talkers are located in the Unmarked Northern quadrant (top left in Fig. 4), and the Mid-Atlantic talkers are located in the Marked Northern quadrant (bottom left in Fig. 4). Although descriptively the Northern dialect is linguistically marked due to the Northern Cities Chain Shift, the perception of the NCCS by Midwestern listeners is more variable. Some listeners, particularly in the Northern dialect region, do not appear to perceive the shifted Northern vowels as distinct from the unmarked vowels produced by Midland talkers (Clopper, 2004; Niedzielski, 1999). The result of this variation in perception is that the Northern talkers appear primarily in the Unmarked Northern quadrant of the space. The two dimensions revealed by the multidimensional scaling analysis are also consistent with the recent findings obtained using a free classification task (Clopper, 2004). These studies provide converging evidence for two primary dimensions of perceptual dialect similarity: geography and markedness.

IV. GENERAL DISCUSSION

The results of experiments 1 and 2 confirm that naive listeners can make explicit judgments about both the identity and the similarity of regional varieties of American English. Overall performance in the four-alternative forced-choice categorization experiment was approximately 42%, which is statistically above chance. In the paired comparison similarity ratings task, the listeners assigned higher similarity ratings to pairs of talkers from the same dialect region than to pairs from different dialect regions. The results of the multidimensional scaling analysis on the similarity ratings were interpretable in terms of the observed linguistic attributes of the four dialects. The two dimensions served to divide the similarity space into four quadrants that approximately corresponded to the four dialects.

An examination of the response biases in the four-alternative forced-choice task in experiment 1 revealed a positive bias for General American responses, but no strong negative biases for the remaining three response alternatives. In our earlier study (Clopper, 2004), we found a positive response bias for the Midland, but negative response biases for New England and West. We argued that the listeners' lack of personal experience with New England and Western talkers led to these negative response biases, whereas familiarity with Midland speech and the location of the listeners in the Midland dialect region at the time of the experiment contributed to a positive response bias (Clopper, 2004). The positive response bias for General American in the current experiment can be attributed to similar factors. In addition, the listeners may have chosen this category more often simply because it covered a larger geographic area than the other categories. Regardless of the factors involved, however, while Midland in the six-alternative task and General American in the four-alternative task were treated as defaults and thus were selected as a response more often than the other response alternatives, negative response biases for less familiar dialects such as New England and West were eliminated in the four-alternative task. Further research is needed with listeners from other geographic areas to determine how the listener's region of origin affects the perception of subvarieties of General American English.

The emergence of gender as a statistically significant factor in the overall perceptual ratings merits discussion. Listeners were not given any explicit instructions about the role that talker gender should play in their similarity judgments. However, gender was found to interact with dialect in an additive manner, so that ratings on same-gender pairs were consistently higher overall than ratings on different-gender pairs, regardless of the dialects of the two talkers. The finding that gender interacts with dialect perception is not too surprising, however. First, gender and dialect are known to interact in speech production. For example, women tend to be more advanced in adopting vowel shifts such as the Northern Cities Chain Shift (e.g., Eckert, 1989). We therefore might expect listeners to be aware of those within-dialect gender differences and that their responses might reflect that awareness. Second,

gender has repeatedly been shown to interact with the perception of the linguistic properties of speech. For example, Strand (1999) found that both talker gender and gender prototypicality affected the perception of /s/ and /ʃ/. And, Mullennix and Pisoni (1990) found indirect evidence of the effects of talker gender on speech perception in a Garner speeded classification task. They reported that variation in talker gender interfered with listeners' performance in a phoneme classification task and led to slower response times than when talker gender was held constant across trials. Taken together, these studies all suggest that talker gender is very difficult to selectively ignore when listeners are asked to make explicit judgments about other aspects of the speech signal.

V. CONCLUSIONS

The results of the current study demonstrate that naive listeners can make explicit judgments about the similarity of pairs of talkers based on regional dialect. These findings are consistent with previous research on the perception of indexical properties of speech which has consistently found that both experienced and naive listeners can make explicit judgments about properties of unfamiliar talker's voices, including voice quality (Kreiman *et al.*, 1994; Remez *et al.*, 2004; Walden *et al.*, 1978), gender (Lass *et al.*, 1976), and sexual orientation (Avery and Liss, 1996; Gaudio, 1994), and suggest that social group properties such as regional, ethnic, and social dialect are also integral components of speech perception and spoken language processing. Dialect variation is a significant source of variability in speech and naive listeners can make explicit judgments about this source of variation in the speech of unfamiliar talkers.

Acknowledgments

This work was supported by NIH NIDCD T32 Training Grant No. DC00012 and NIH NIDCD R01 Research Grant No. DC00111 to Indiana University.

References

- Abercrombie, D. *Elements of General Phonetics*. Aldine; Chicago: 1967.
- Avery JD, Liss JM. Acoustic characteristics of less-masculine-sounding speech. *J Acoust Soc Am*. 1996; 99:3738–3748. [PubMed: 8655805]
- Clopper, CG. PhD dissertation. Indiana University; 2004. Linguistic experience and the perceptual classification of dialect variation.
- Clopper CG, Pisoni DB. The Nationwide Speech Project: A new corpus of American English dialects. *Speech Commun.* (unpublished).
- Clopper CG, Conrey BL, Pisoni DB. Effects of talker gender on dialect categorization. *J Lang Soc Psychol*. 2005a; 24:182–206. [PubMed: 21423866]
- Clopper CG, Pisoni DB, de Jong K. Acoustic characteristics of the vowel systems of six regional varieties of American English. *J Acoust Soc Am*. 2005b; 118:1661–1676. [PubMed: 16240825]
- Eckert P. The whole woman: Sex and gender differences in variation. *Lang Var Change*. 1989; 1:245–267.
- Gaudio RP. Sounding gay: Pitch properties in the speech of gay and straight men. *Am Speech*. 1994; 69:30–57.
- Goldinger SD. Words and voices: Episodic traces in spoken word identification and recognition memory. *J Exp Psychol Learn Mem Cogn*. 1996; 22:1166–1183. [PubMed: 8926483]
- Kalikow DN, Stevens KN, Elliott LL. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J Acoust Soc Am*. 1977; 61:1337–1351. [PubMed: 881487]
- Kreiman J, Gerratt BR, Berke GS. The multidimensional nature of pathologic vocal quality. *J Acoust Soc Am*. 1994; 96:1291–1302. [PubMed: 7962996]

- Labov, W. Sociolinguistic Patterns. Univ. of Pennsylvania; Philadelphia: 1972.
- Labov, W. The three dialects of English. In: Linn, MD., editor. Handbook of Dialects and Language Variation. Academic; San Diego: 1998. p. 39-81.
- Lass NJ, Hughes KR, Bowyer MD, Waters LT, Bourne VT. Speaker sex identification from voiced, whispered, and filtered isolated vowels. *J Acoust Soc Am*. 1976; 59:675–678. [PubMed: 1254794]
- Mase, Y. Dialect consciousness and dialect divisions: Examples in the Nagano-Gifu boundary region. In: Preston, DR., editor. Handbook of Perceptual Dialectology. John Benjamins; Amsterdam: 1999. p. 71-99.
- Miller GA. What is information measurement? *Am Psychol*. 1953; 8:3–11.
- Miller GA, Nicely PE. An analysis of perceptual confusions among some English consonants. *J Acoust Soc Am*. 1955; 27:338–352.
- Mullennix JW, Pisoni DB. Stimulus variability and processing dependencies in speech perception. *Percept Psychophys*. 1990; 47:379–390. [PubMed: 2345691]
- Nerbonne J, Heeringa W. Computational comparison and classification of dialects. *Dialectol Geolinguist*. 2001; 9:69–83.
- Nerbonne, J.; Heeringa, W.; vanden Hout, E.; van der Kooi, P.; Otten, S.; van de Vis, W. Phonetic distance between Dutch dialects. In: Durieuz, G.; Daelemans, W.; Gillis, S., editors. Papers from the 6th CLIN Meeting. Univ. of Antwerp, Center for Dutch Language and Speech; Antwerp: 1996. p. 185-202.
- Niedzielski N. The effect of social information on the perception of sociolinguistic variables. *J Lang Soc Psychol*. 1999; 18:62–85.
- Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. *Percept Psychophys*. 1998; 60:355–376. [PubMed: 9599989]
- Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. *Psychol Sci*. 1994; 5:42–46. [PubMed: 21526138]
- Remez RE, Wissig SC, Ferro DF, Liberman K, Landau C. A search for listener differences in the perception of talker identity. *J Acoust Soc Am*. 2004; 116:2544.
- Sternberg, S. Discovering mental processing stages: The method of additive factors. In: Scarborough, D.; Sternberg, S., editors. An Invitation to Cognitive Science: Methods, Models, and Conceptual Issues. MIT; Cambridge, MA: 1998. p. 703-864.
- Strand EA. Uncovering the role of gender stereotypes in speech perception. *J Lang Soc Psychol*. 1999; 18:86–100.
- Thomas, ER. An Acoustic Analysis of Vowel Variation in New World English. Duke U. P; Durham, NC: 2001.
- Walden BE, Montgomery AA, Gibeily GJ, Prosek RA, Schwartz DM. Correlates of psychological dimensions in talker similarity. *J Speech Hear Res*. 1978; 21:265–275. [PubMed: 703276]

APPENDIX: STIMULUS MATERIALS

Key: Mid-Atlantic (AT), North (NO), South (SO), Midland (MI), New England (NE), West (WE), General American (GA).

Talker	List A	List B
AT1	He rode off in a cloud of dust.	Ruth poured herself a cup of tea.
AT2	The car drove off the steep cliff.	Ruth had a necklace of glass beads.
AT3	Paul hit the water with a splash.	My son has a dog for a pet.
AT5	The cow gave birth to a calf.	Banks keep their money in a vault.
AT6	The glass had a chip on the rim.	I ate a piece of chocolate fudge.
AT7	The heavy rains caused a flood.	Cut the meat into small chunks.

Talker	List A	List B
AT9	Throw out all this useless junk.	Watermelons have lots of seeds.
A18	Kill the bugs with this spray.	Please wipe your feet on the mat.
NO0	Tighten the belt by a notch.	That job was an easy task.
NO2	The cabin was made of logs.	They tracked the lion to his den.
NO3	Peter dropped in for a brief chat.	Wash the floor with a mop.
NO4	The cut on his knee formed a scab.	Raise the flag up the pole.
NO5	The shepherd watched his flock of sheep.	The story had a clever plot.
NO6	The swimmer's leg got a bad cramp.	Paul took a bath in the tub.
NO8	The flashlight casts a bright beam.	The mouse was caught in the trap.
NO9	Bob was cut by the jackknife's blade.	Paul was arrested by the cops.
SO1	A bicycle has two wheels.	The landlord raised the rent.
S10	We swam at the beach at high tide.	The guests were welcomed by the host.
SO2	My jaw aches when I chew gum.	The sick child swallowed the pill.
S22	Spread some butter on your bread.	Playing checkers can be fun.
SO5	Get the bread and cut me a slice.	The scarf was made of shiny silk.
SO6	The judge is sitting on the bench.	We camped out in our tent.
SO7	Greet the heroes with loud cheers.	The bride wore a white gown.
SO8	For dessert he had apple pie.	She cooked him a hearty meal.
M12 (GA)	The lion gave an angry roar.	The chicken pecked corn with its beak.
MI3 (GA)	He was scared out of his wits.	The detectives searched for a clue.
MI8 15(GA)	He's employed by a large firm.	The bloodhound followed the trail.
NE1 (GA)	The shepherds guarded their flock.	A spoiled child is a brat.
NE7 (GA)	The doctor prescribed the drug.	To open the jar, twist the lid.
NE8 (GA)	Unlock the door and turn the knob.	Keep your broken arm in a sling.
WE2 (GA)	She shortened the hem of her skirt.	Our seats were in the second row.
WE8 (GA)	The witness took a solemn oath.	The super highway has six lanes.

FIG. 1.
Response alternatives in the four-alternative forced-choice categorization task. In the colored map that the participants saw, Mid-Atlantic was magenta, North was red, South was orange, and General American was green.

FIG. 2. Percent correct responses in the four-alternative forced-choice categorization task for each of the four talker dialect groups, collapsed across experimental block. Error bars indicate standard error. Chance performance (25%) is indicated by the dashed line.

FIG. 3. Mean similarity ratings for same-dialect and different-dialect trials for each of the four listener groups, collapsed across experimental block and stimulus list. Error bars indicate standard error.

FIG. 4.

Multidimensional scaling solution. Filled symbols represent male talkers and open symbols represent female talkers.

TABLE I

Mean normalized subject weights from the INDSCAL analysis for each of the four listener groups and each of the two stimulus lists for each of the two dimensions.

	Dimension 1 geography	Dimension 2 markedness
Mobile North	0.50	0.50
Mobile Midland	0.49	0.51
Non-Mobile North	0.51	0.49
Non-Mobile Midland	0.51	0.49
List A	0.48	0.52
List B	0.53	0.47